Overview of the TREC 2014 Federated Web Search Track

Thomas Demeester[1], Dolf Trieschnigg[2], Dong Nguyen[2], Ke Zhou[3], Djoerd Hiemstra[2]
[1] Ghent University - iMinds, Belgium
[2] University of Twente, The Netherlands
[3] Yahoo Labs London, United Kingdom
tdmeeste@intec.ugent.be, {d.trieschnigg, d.nguyen}@utwente.nl,
kezhou@yahoo-inc.com, d.hiemstra@utwente.nl

ABSTRACT

The TREC Federated Web Search track facilitates research on federated web search, by providing a large realistic data collection sampled from a multitude of online search engines. The FedWeb 2013 Resource Selection and Results Merging tasks are again included in FedWeb 2014, and we additionally introduced the task of vertical selection. Other new aspects are the required link between the Resource Selection and Results Merging tasks, and the importance of diversity in the merged results. After an overview of the new data collection and relevance judgments, the individual participants' results for the tasks are introduced, analyzed, and compared.

1. INTRODUCTION

When Sergey Brin and Larry Page wrote their seminal "The Anatomy of a Large-Scale Hypertextual Web Search Engine" [3] they added an appendix about the scalibility of Google in which they argued that its scalability is limited by their choice for a single, centralized index. While these limitations would decrease over time, following Moore's law, a truly scalable solution would require a drastic redesign. They write the following:

> "Of course a distributed systems like Gloss or Harvest will often be the most efficient and elegant technical solution for indexing, but it seems difficult to convince the world to use these systems because of the high administration costs of setting up large numbers of installations. Of course, it is quite likely that reducing the administration cost drastically is possible. If that happens, and everyone starts running a distributed indexing system, searching would certainly improve drastically." (Brin and Page 1998 [3])

When we started to crawl results from independent web search engines of all kinds, we hoped it would inspire researchers to come up with elegant and efficient solutions to distributed search. However, the crawl can be used for many other research goals as well, including scenarios that resemble the aggregated search approaches implemented by most general web search engines today.

The TREC federated web search track provides a test collection consisting of search result pages of 149 internet search engines. The track aims to answer research questions like: "What is the best search engine for this query?", "What is the best medium, topic or genre, for this query?" and "How do I combine the search results of a selection of the search engines into one coherent ranked list?" The research questions are addressed in the following three tasks: Resource Selection, Vertical Selection, and Results Merging:

Task 1: Resource Selection
The goal of resource selection is to select the right resources (search engines) from a large number of independent search engines given a query. Participants have to rank the given 149 search engines for each test topic without having access to the corresponding search results. The FedWeb 2014 collection contains search result pages for many other queries, as well as the HTML of the corresponding web pages. These data could be used by the participants to build resource descriptions. Participants may also use external sources such as Wikipedia, ODP, or WordNet.

Task 2: Vertical Selection
The goal of vertical selection is to classify each query into a fixed set of 24 verticals, i.e. content dedicated to either a topic (e.g. "finance"), a media type (e.g. "images") or a genre (e.g. "news"). Each vertical contains several resources, for example, the "image" vertical contains resources such as Flickr and Picasa. With this task, we aim to encourage vertical (domain) modeling from the participants.

Task 3: Results Merging
The goal of results merging is to combine the results of several search engines into a single ranked list. After the deadline for Task 1 passed, the participants were given the search result pages of 149 search engines for the test topics. The result pages include titles, result snippets, hyperlinks, and possibly thumbnail images, all of which were used by participants for reranking and merging.

The official FedWeb track guidelines can be found online[1]. This overview paper is organized as follows: Section 2 describes the FedWeb 2014 collection; Section 3 describes the process of gathering relevance judgements for the track; Section 4 presents our online system for validation and preliminary evaluation of runs. Sections 5, 6 and 7 describe the results for the vertical selection task, the resource selection

[1] `http://snipdex.org/fedweb`

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **NOV 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE **Overview of the TREC 2014 Federated Web Search Track**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Ghent University -iMinds,Belgium,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

The TREC Federated Web Search track facilitates research on federated web search, by providing a large realistic data collection sampled from a multitude of online search engines. The FedWeb 2013 Resource Selection and Results Merging tasks are again included in FedWeb 2014, and we additionally introduced the task of vertical selection. Other new aspects are the required link between the Resource Selection and Results Merging tasks, and the importance of diversity in the merged results. After an overview of the new data collection and relevance judgments, the individual participants??? results for the tasks are introduced, analyzed, and compared.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

Vertical	# Resources
Academic	17
Video	11
Photo/Pictures	11
Health	11
Shopping	10
News	10
General	8
Encyclopedia	8
Sports	7
Kids	7
Q&A	6
Games	6
Tech	5
Recipes	5
Jobs	5
Blogs	4
Software	3
Social	3
Entertainment	3
Travel	2
Jokes	2
Books	2
Audio	2
Local	1

Table 1: Vertical statistics

task and the results merging task, respectively; Section 8 gives a summary of this year's track main findings.

2. FEDWEB 2014 COLLECTION

Similar to last year the collection for the FedWeb track consisted of a *sample* crawl and a *topic* crawl for a large number of online search engines. The *sample* crawl consists of sampled search engine results (i.e. the snippets from the first 10 results) and downloads of the pages these snippets refer to. The snippets and pages can be used to create a resource description for each search engine, to support vertical and resource selection. The *topic* crawl is used for evaluation and consists of only the snippets for a number of topic queries. In contrast to last year, in which also the pages of the topic queries were available, we provided only the snippets of the topics to make the tasks more realistic.

Where possible we reused the list of search engines from the 2013 track, ending up with a list of 149 search engines which were still available for crawling. We doubled the number of sample queries to 4000, to allow for more precise resource descriptions. Similar to last year the first set of 2000 queries were based on single words sampled from different frequency bins from the vocabulary of the ClueWeb09-A collection. These correspond to the sample queries issued in 2013. The second set of 2000 queries is different for each engine and consists of random words sampled from the language model obtained from the first 2000 snippets.

Table 1 lists the number of resources (search engines) per vertical. Appendix A lists the engines used this year.

3. RELEVANCE ASSESSMENTS

In this section, we describe how the test topics were chosen and how the relevance judgments were organized. We also visualize the distribution of relevant documents over the different test topics, and over the various verticals.

3.1 Test Topics

We started from the 506 topics gathered for FedWeb 2013 [5], leaving out the 200 topics provided to the FedWeb 2013 participants. From the remaining 306 topics, we selected 75 topics as follows. We first assigned labels of the most likely vertical intents to each of the topics (based on intuition and query descriptions). We then manually selected these 75 topics such, that most of the topics would potentially target other verticals than just general web search engines, where even the smallest verticals had at least one dedicated topic (e.g., Jokes, or Games), and with more emphasis on the larger verticals (see Appendix A). The pages from all resources were entirely judged for 60 topics, randomly chosen among the 75 selected ones. The first 10 fully annotated topics were used for the online evaluation system (see Section 4), and the remaining 50 are the actual test topics (see Appendix B).

For the previous (2013) edition of the track, we had the top 3 snippets from each search engine for each of the candidate topics judged first, on which we based the choice of evaluation topics, and which provided the starting point for writing out the narratives providing the annotation context. This year, we decided not to do any snippet judgments, and instead, to spend our resources on judging 10 extra topics. We manually created the narratives by quickly going through the results, and in consultation with the assessors. An example of one of the test topics is given below, with the query terms, description, and narrative, which were all shown to the assessors. Each topic was judged by a single assessor, in a random order, where we had contributions from 10 hired assessors. The assessors are all students in various fields, such that we had the liberty of assigning specialized queries to specialized assessors. For example, the topic given below was entirely judged by a medical student.

```
<topic id="7215">
  <query>squamous cell carcinoma</query>
  <description>You are looking for information about
    Squamous Cell Carcinoma (skin cancer).
  </description>
  <narrative>You have been diagnosed with squamous cell
    carcinoma. You are looking for information, including
    treatments, prognosis, etc. Given your medical
    background (you are a doctor), you want to  search
    the existing literature in depth, and are most
    interested in scientific results.
  </narrative>
</topic>
```

3.2 Relevance Levels

The same graded relevance levels were used as in the FedWeb 2013 edition, taken over from the TREC Web Track[2]: Non (not relevant), Rel (minimal relevance), HRel (highly relevant), Key (top relevance), and Nav (navigational). Based on the User Disagreement Model (UDM), introduced in [4],

[2] http://research.microsoft.com/en-us/projects/trec-web-2013/

the following weights are assigned to these relevance levels:

$$\begin{aligned} w_{\text{Non}} &= 0.0 \\ w_{\text{Rel}} &= 0.158 \\ w_{\text{HRel}} &= 0.546 \\ w_{\text{Key}} &= 1.0 \\ w_{\text{Nav}} &= 1.0 \end{aligned}$$

These were estimated from a set of double annotations for the FedWeb 2013 collection, which has, by construction, comparable properties to the FedWeb 2014 dataset.

For evaluating the quality of a set of 10 results as returned by the resources in response to a test topic, we use the relevance weights listed above to calculate the Graded Precision (introduced by [11] as the generalized precision). This measure amounts to the sum of the relevance weights associated with each of the results, divided by 10 (also for resources that returned less than 10 results).

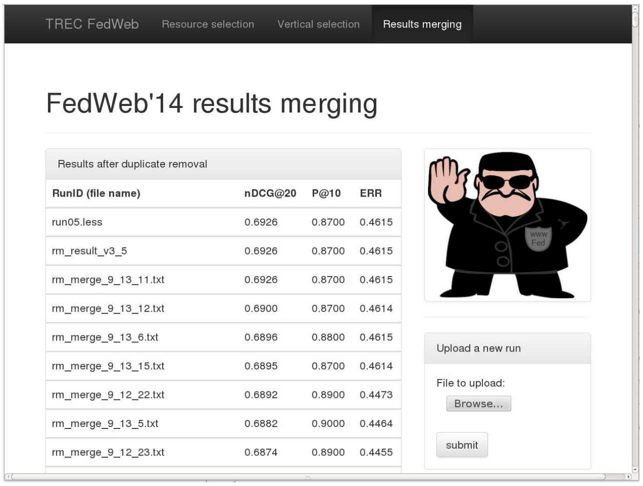
We now provide some insights into how the most relevant documents are distributed, depending on the test topics and among the different verticals. Fig. 1 shows, for each test topic, the highest graded precision as found among all resources. The figure can thus be interpreted as a ranking of the topics from 'easy' to 'difficult', with respect to the set of resources in the FedWeb 2014 system. For example, for the leftmost topic 7252, one resource managed to return 10 Key results (not taking into account duplicate results). The query *welch corgi* targeted broad information, including pictures and videos, on Welsh corgi dogs. For the rightmost topic 7222, no Key results were returned, although a number of HRel results were. The query *route 666* appeared to be rather ambiguous, and the narrative specified a specific need only (reviews/summaries of the movie).

Next, we selected for each topic the best resource (i.e., with highest graded precision) within each of the verticals, and created a boxplot by aggregating over the verticals. The result is shown in Fig. 2. We see that the best resource (depending on the queries) from the General search engines achieves the highest number of relevant results (and/or the results with the highest levels of relevance), followed by the Blogs, Kids, and Video verticals.

4. PRELIMINARY ONLINE EVALUATION

During the last couple of weeks before the submission deadline for the different tasks, we opened up an online platform where participants could test their systems under preparation. By submitting a preliminary run to this system, the runs were validated by checking if they adhere to the TREC format, and the main evaluation metrics were returned. The evaluation metrics returned were based on 10 test queries, i.e., as described above, those 10 that were fully annotated but not used for the actual evaluation. Figure 3 shows a screenshot of the online system.

Multiple participants indeed used this system, and we kept track of the different submitted runs. More than 500 runs were validated and tested online before the official submission deadline. Figure 4 shows the main evaluation metrics (F1 for Vertical Selection, and nDCG@20 for both Resource Selection and Results Merging) for the valid runs among the online trial submissions. These metrics are the results with respect to the 50 evaluation topics, not including the 10 test topics for which the participants received the intermediate results (and towards which their systems might have been tuned). We did not try to link trial runs to specific participants, although we noticed that the same team often submitted consecutive runs to the system, either for a range of different techniques, or maybe to determine suitable values for model hyperparameters. For the Vertical Selection task, there is an overall increase in effectiveness of the systems, although the last runs seem to perform worse. For the Resource Selection task, the best run was found early on in the chronological order. For the Results Merging tasks more than half of the runs perform almost equally well, around nDCG@20≈0.3, although few runs perform better, which might be explained by the fact that participants over-trained their systems on the 10 test queries of the online system.



FedWeb'14 results merging

Results after duplicate removal

RunID (file name)	nDCG@20	P@10	ERR
run05.less	0.6926	0.8700	0.4615
rm_result_v3_5	0.6926	0.8700	0.4615
rm_merge_9_13_11.txt	0.6926	0.8700	0.4615
rm_merge_9_13_12.txt	0.6900	0.8700	0.4614
rm_merge_9_13_6.txt	0.6896	0.8800	0.4615
rm_merge_9_13_15.txt	0.6895	0.8700	0.4614
rm_merge_9_12_22.txt	0.6892	0.8900	0.4473
rm_merge_9_13_5.txt	0.6882	0.9000	0.4464
rm_merge_9_12_23.txt	0.6874	0.8900	0.4455

Upload a new run

File to upload:

Browse...

submit

Figure 3: Screen shot of the online evaluation system.

5. VERTICAL SELECTION

5.1 Evaluation

We report the precision, recall and F-measure (primary metric) of the submitted vertical selection runs in Table 2. The primary vertical selection evaluation metric is the F-measure (based on our own implementation). The methodology of how we obtain the vertical relevance can be referred to the (GMR + II) approach described in [18]. Basically, the relevance of a vertical for a given query is determined by the best performing resource (search engine) within this vertical. More specifically, the relevance is represented by the maximum graded precision of its resources. For the final evaluation, the binary relevance of a vertical is determined by a threshold: a vertical for which the maximum graded precision is 0.5 or higher, is considered relevant. This threshold was determined based on data analyses, such that for most queries there is a small set of relevant verticals. If for a given query, no verticals have exceeded this threshold, we use the top-1 vertical with the maximal relevance as the relevant vertical.

5.2 Analysis

Seven teams participated in the vertical selection task, with a total of 32 system runs. The four best performing runs based on the F-measure (`ICTNETVS07`, `esevsru`, `esevs` and `ICTNETVS02`) were submitted by East China Normal

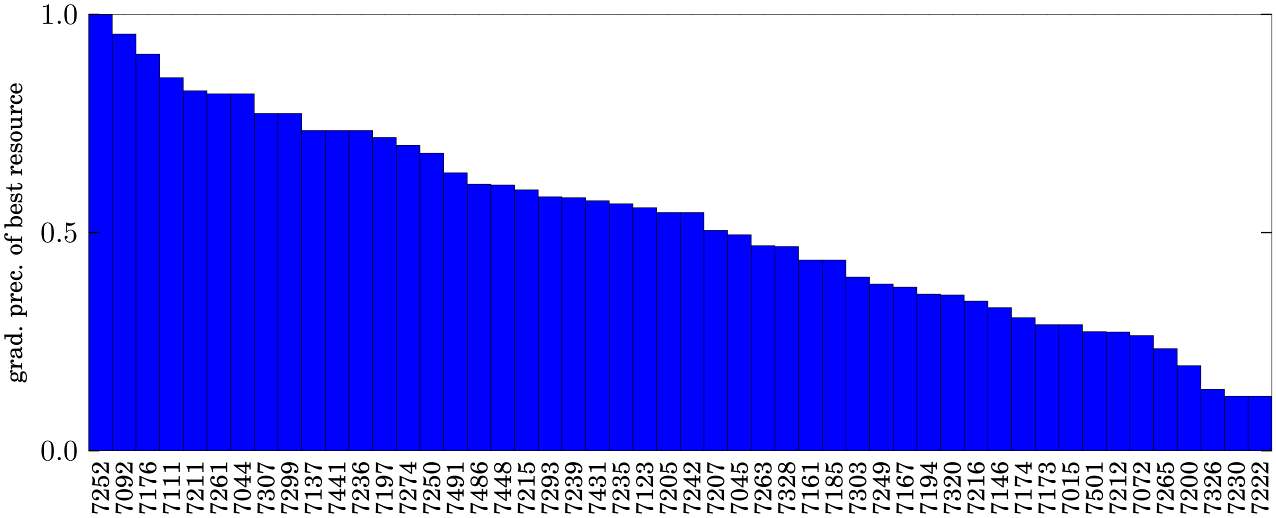


Figure 1: Graded relevance of the best resource per topic, for all 50 test topics.

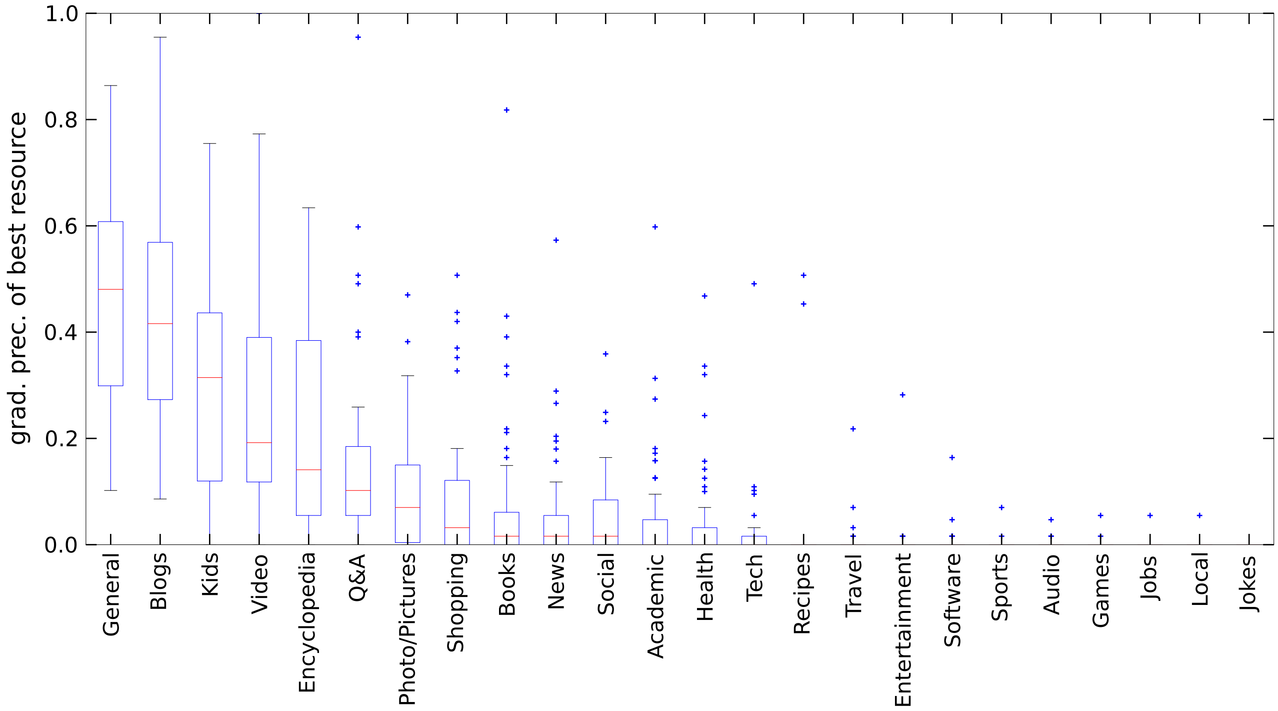


Figure 2: Highest graded relevance among all resources within a vertical, over all 50 test topics.

University (ECNUCS) and Chinese Academy of Sciences, Inst. of Computing Technology (ICTNET). Interestingly, the top-1 run (`ICTNETVS07`) utilized the documents as the sole source of evidence in selecting verticals while all the other top runs exploited external resources, such as Google API, WEKA or KDD 2005 data.

5.3 Participant Approaches

Chinese Academy of Sciences (ICTNET) [8]

For the task of Vertical Selection, ICTNET submitted a number of high-scoring runs, including the overall best performing run (`ICTNETVS07`). Several strategies were proposed. For `ICTNETVS1`, they calculated a term frequency based similarity score between queries and verticals. They also explored using random forest classification to score verticals

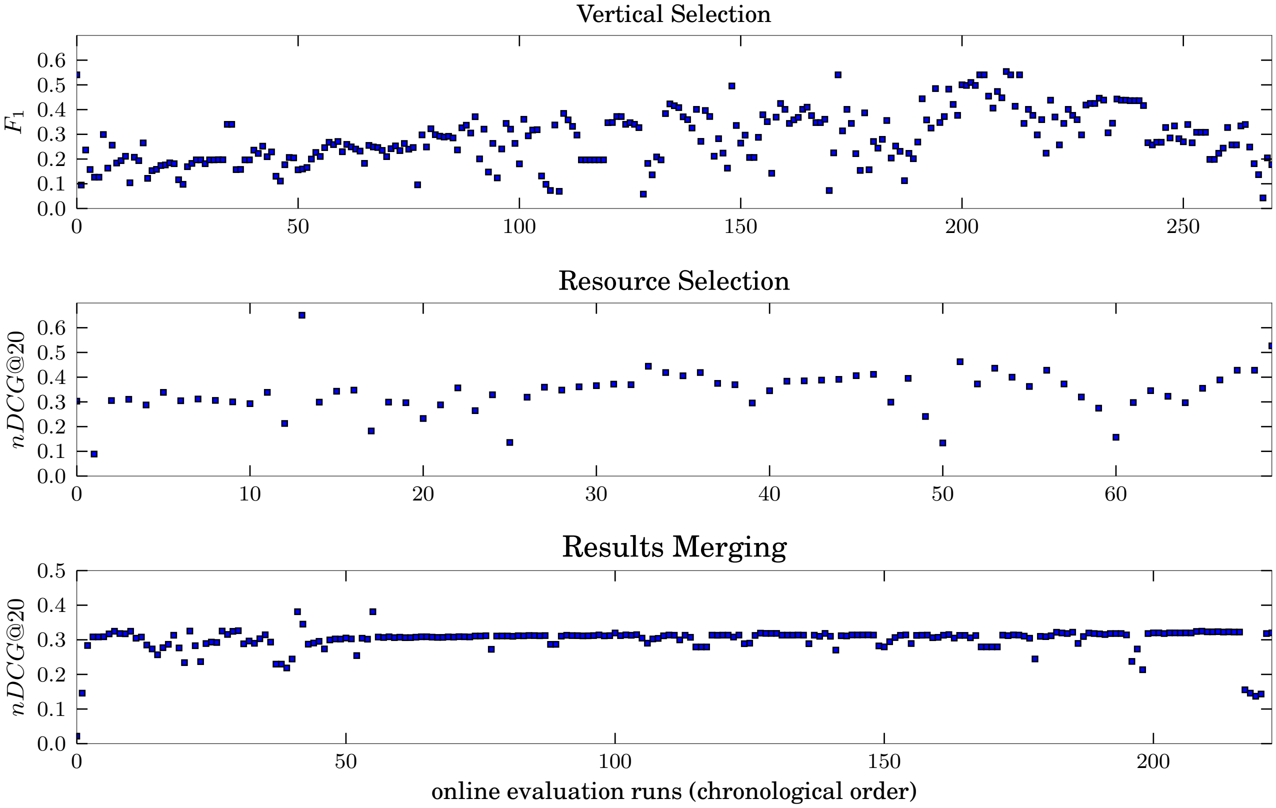


Figure 4: Main metrics per task, for the trial runs, in the order as submitted to the online evaluation system.

(run ICTNETVS02), whereby expanded query representations based on results from the Google Custom Search API were used. They further used a model to calculate the similarity between a vertical (represented by a small portion of the available documents) and the expanded query representation, based on Latent Semantic Indexing (LSI) to score verticals (with run ICTNETVS03). They also submitted a number of runs with variations and/or combinations of these methods (ICTNETVS04, ICTNETVS05, ICTNETVS06). For ICTNETVS07, the best run for this task, they used a borda fuse combination of 3 methods, based on frequent term ranks in the given documents.

East China Normal University (ECNUCS) [10]

East China Normal University introduces the Search Engine Impact Factor (SEIF), a query-independent measure of a search engine's impact, estimated in two different ways: 1) using external data from comScore, a company providing marketing data and analytics to web pages of many enterprises and publishers; and 2) using the TREC 2013 dataset and its relevance judgments. Their best vertical selection run (esevsru) was the overall second best. It combines three methods: 1) matching WordNet synonyms for queries and verticals, 2) training a classifier on the KDD Cup 2005 Internet user search query classification dataset [12], and 3) the search engine categories provided by FedWeb. Their other runs are based on a single, or combination of two of the above methods.

University of Delaware (udel) [1]

Both submitted runs are based on the resource selection run udelftrsbs, whereby the baseline udelftvql ranks verticals according to the number of resources in the corresponding resource selection run, and for udelftvqlR some vertical-specific rule-based modifications were done (e.g., to require the presence of interrogative words for the Q&A vertical), resulting in a significant increase in the F-measure.

Drexel University (dragon) [16]

Drexel University's approach for vertical selection was based on their resource selection methodology. To select only a subset of verticals from the vertical ranking, they set a fixed cut-off threshold 0.01 on the normalized vertical score. This fixed threshold also resulted in high recall and low precision while the CRCS approach (drexelVS1) performed the best.

University of Stavanger (NTNUiS) [2]

Their vertical selection runs were directly based on their resource selection runs. In particular, they applied a threshold on the relevance scores of the individual resources and selected all verticals containing a resource that passed a threshold. The NTNUiSvs2 run, based on their best performing resource selection run, performed best.

Task 2: Vertical Selection					
Group ID	Run ID	Precision	Recall	**F-measure**	Resources Used
ECNUCS	ekwma	0.054	0.120	0.069	snippets, wordnet
	esevs	0.398	0.586	0.438	snippets, trec 2013 dataset, kdd 2005
	esevsru	0.388	0.598	0.440	snippets, trec 2013 dataset, kdd 2005
	esvru	0.276	0.439	0.297	snippets, kdd 2005, google search
	svmtrain	0.338	0.425	0.338	snippets, kdd 2005, google search
ICTNET	ICTNETVS02	0.292	0.790	0.401	documents, Google API, WEKA
	ICTNETVS03	0.276	0.410	0.298	snippets, documents, Google API, NLTK, GENSIM
	ICTNETVS04	0.427	0.392	0.377	snippets, documents, Google API, NLTK, GENSIM, WEKA
	ICTNETVS05	0.423	0.365	0.359	snippets, documents, Google API, NLTK, GENSIM, WEKA
	ICTNETVS06	0.258	0.673	0.344	documents, Google API, WEKA
	ICTNETVS07	0.591	0.545	0.496	documents
	ICTNETVS1	0.230	0.638	0.299	snippets, documents
NTNUiS	NTNUiSvs2	0.157	0.406	0.205	snippets, documents
	NTNUiSvs3	0.145	0.281	0.177	snippets, documents
ULugano	ULuganoCL2V	0.117	0.983	0.197	documents, SentiWordNet Lexicon
	ULuganoDFRV	0.117	0.983	0.197	documents
	ULuganoDL2V	0.117	0.983	0.197	documents, SentiWordNet Lexicon
UPD	UPDFW14v0knm	0.076	1.000	0.138	documents
	UPDFW14v0nnm	0.076	1.000	0.138	documents
	UPDFW14v0pnm	0.076	1.000	0.138	documents
	UPDFW14v1knm	0.076	1.000	0.138	documents
	UPDFW14v1nnm	0.076	1.000	0.138	documents
	UPDFW14v1pnm	0.076	1.000	0.138	documents
dragon	drexelVS1	0.240	0.506	0.284	documents
	drexelVS2	0.159	0.824	0.233	documents
	drexelVS3	0.134	0.960	0.212	documents
	drexelVS4	0.134	0.960	0.212	documents
	drexelVS5	0.163	0.824	0.244	documents
	drexelVS6	0.171	0.729	0.251	documents
	drexelVS7	0.189	0.732	0.271	documents
udel	udelftvql	0.167	0.852	0.257	documents
	udelftvqlR	0.236	0.680	0.328	documents

Table 2: Results for the Vertical Selection task.

University of Lugano (ULugano) [7]

The vertical selection runs they submitted were simply a direct derivation from their resource selection runs. Basically, for each of the resource selection run, they simply aggregated the resource selection scores of the resources within each vertical and did not set any thresholds on the number of selected verticals. Therefore, this resulted in the high recall and low precision of all their vertical selection runs.

University of Padova (UPD) [6]

The University of Padova's participation aimed at the investigation of the effectiveness of the TWF.IRF weighting algorithm in a Federated Web search setting. TWF.IRF, Term Weighted Frequency times Inverse Resource Frequency, is a recursive weighting scheme originally proposed for hybrid hierarchical peer-to-peer networks. The University of Padova looked into the influence of stemming and stopwords. Their results indicate that stemming has no significant effect on TWF.IRF effectiveness, and that overall the TWF.IRF approach is not highly effective for vertical selection.

6. RESOURCE SELECTION

6.1 Evaluation

We report the nDCG@20 (primary metric), nDCG@10, nP@1 and nP@5 of the submitted resource selection runs in Table 3. The primary evaluation metric is nDCG@20 (using the implementation of `ndcg_cut.20` in `trec_eval`). The relevance of a resource for a given query is obtained by calculating the graded precision (see Section 3.2) on the top 10 results. These values are used as the nDCG gain values, for convenience with `trec_eval` scaled by a factor of 1000. Thus, this metric takes the ranking of resources into account and the graded relevance of the documents in the top 10 of each resource, but not the ranking of documents *within* the resources.

We also report nP@1 and nP@5 (normalized graded precision at *k=1* and *k=5*). Introduced in the FedWeb 2013 track [5], the normalized graded precision represents the graded precision of the top ranked k resources, normalized by the graded precision of the best possible k resources for the given topic. Compared to nDCG, this metrics ignores the ranking of the resources within the top k. For example, nP@1 denotes the graded precision of the highest ranked resource, divided by the highest graded precision by any of the resources for that topic.

6.2 Analysis

This year, 10 teams participated in the resource selection task, with a total of 44 runs. The four best performing runs based on nDCG@20 (`ecomsvz`, `ecomsv`, `eseif` and `ecomsvt`) were all submitted by East China Normal University (ECNUCS). These runs only make use of result snippets, and their ranking strategies are based for an important part on the Search Engine Impact Factor. In addition, three of these runs (`ecomsvz`, `ecomsv` and `ecomsvt`) make use of external resources (Google Search, data from KDD 2005). Interestingly, their `eseif` run is a static, query-independent ranking based on data from the Fedweb TREC 2013 task. The top 5 resources of their static run are: Yahoo Screen, Yahoo Answers, AOL Video, Kidrex and Ask. The second team, info_ruc, used query extension based on Google, and matched queries with resources, based on a topic model representation, whereby a snippet-based topic model proved consistently better than one based on full web documents.

6.3 Participant Approaches

East China Normal University (ECNUCS) [10]

Their resource selection runs outperform the runs from other participants by a big margin. For their best run (`ecomsvz`), several techniques were combined to score resources for each query. The Search Engine Impact Factor (see ECNUCS' vertical selection submissions) has the biggest contribution to performance improvements, besides the vertical selection results, tf-idf features, and a semantic similarity score. The indivitual contributions from these methods are explored in the other submitted runs.

Renmin University of China (info_ruc) [15]

The team info_ruc used two different LDA topic distributions for its resource selection runs. For the runs `FW14DocsX` (`X`=50, 75, 100), they performed an LDA analysis over the whole set of sampled documents, after which the topic distribution of each resource was determined as the average distribution of its documents. For the runs `FW14SearchX`, they merged all sampled snippets into one big document, and used these to infer LDA topics from. `X` represents the number of topics used. Each query was expanded using the Google Search API, and its topic distribution vector was determined, after which the similarity between the query and resource representation was used to rank resources. The results show that all snippet based runs `FW14SearchX` outperform the sample documents based runs `FW14DocsX`, and resulted in the overall second best set of runs for this task (after the ECNUCS runs). For the snippets, 50 topics were the better choice, against 100 topics for the documents.

Chinese Academy of Sciences (ICTNET) [8]

ICTNET used various approaches for this task. For their first run (`ICTNETRS01`), they used a straightforward IR setup, based on indexing the provided sample documents, to score a resource, thereby giving more weight to higher ranked results. This run performed very low, but augmenting the method with the (highly successful) vertical selection results, resulted in a much better effectiveness (runs `ICTNETRS02` and `ICTNETRS07`). Further runs use a text classification strategy (`ICTNETRS03`) and LSI (`ICTNETRS04`), including the resources' pagerank for the latter. These approaches are similar to the corresponding vertical selection approaches (including the query expansion part). ICTNET's most successful resource selection runs use the LSI model (with pagerank), together with the vertical selection results (`ICTNETRS05` and `ICTNETRS06`).

Drexel University (dragon) [16]

In total 7 runs were submitted and the aim was to evaluate a variety of existing resource selection approaches from the existing literatures, namely ReDDE, ReDDE.top, CRCSLinear, CRCSExp, CiSS, CiSSAprox, SUSHI. All those resource selection approaches are based on the central sampled index (CSI) while the differences of those approaches are how they reward each resource based on the retrieved documents from the CSI. Ultimately, they found that the SUSHI approach (`drexelRS7`) performed the best.

Task 1: Resource Selection						
Group ID	Run ID	**nDCG@20**	nDCG@10	nP@1	nP@5	resources used
ECNUCS	ecomsv	0.700	0.601	0.525	0.579	snippets, Google search, KDD 2005
	ecomsvt	0.626	0.506	0.273	0.491	snippets, Google search, KDD 2005
	ecomsvz	0.712	0.624	0.535	0.604	snippets, Google search, KDD 2005
	eseif	0.651	0.623	0.306	0.546	snippets
	esmimax	0.299	0.261	0.222	0.265	snippets, Google search
	etfidf	0.157	0.113	0.093	0.113	snippets
ICTNET	ICTNETRS01	0.268	0.226	0.163	0.193	documents
	ICTNETRS02	0.365	0.322	0.289	0.324	documents, Google API, NLTK, GENSIM
	ICTNETRS03	0.400	0.340	0.160	0.351	documents, Google API, NLTK, GENSIM, WEKA
	ICTNETRS04	0.362	0.306	0.116	0.290	documents, Google API, NLTK, GENSIM
	ICTNETRS05	0.436	0.391	0.489	0.377	documents, Google API, NLTK, GENSIM
	ICTNETRS06	0.428	0.372	0.521	0.345	documents, Google API, NLTK, GENSIM
	ICTNETRS07	0.373	0.334	0.267	0.334	documents, Google API, NLTK, GENSIM
NTNUiS	NTNUiSrs1	0.306	0.225	0.148	0.195	documents
	NTNUiSrs2	0.348	0.281	0.206	0.257	snippets, documents
	NTNUiSrs3	0.248	0.205	0.202	0.189	snippets, documents
ULugano	ULuganoColL2	0.297	0.189	0.148	0.158	documents, SentiWordNet
	ULuganoDFR	0.304	0.193	0.137	0.164	documents
	ULuganoDocL2	0.301	0.193	0.137	0.160	documents, SentiWordNet
UPD	UPDFW14r1ksm	0.292	0.209	0.148	0.180	documents
	UPDFW14tiknm	0.278	0.209	0.118	0.191	documents
	UPDFW14tiksm	0.310	0.223	0.126	0.188	documents
	UPDFW14tinnm	0.281	0.212	0.134	0.201	snippets, documents
	UPDFW14tinsm	0.306	0.221	0.153	0.197	documents
	UPDFW14tipnm	0.280	0.212	0.115	0.191	snippets, documents
	UPDFW14tipsm	0.311	0.226	0.123	0.187	documents
dragon	drexelRS1	0.389	0.348	0.222	0.318	documents
	drexelRS2	0.328	0.227	0.125	0.180	documents
	drexelRS3	0.333	0.229	0.125	0.179	documents
	drexelRS4	0.333	0.229	0.125	0.180	documents
	drexelRS5	0.342	0.241	0.135	0.211	documents
	drexelRS6	0.382	0.284	0.201	0.250	documents
	drexelRS7	0.422	0.359	0.293	0.314	documents
info_ruc	FW14Docs100	0.444	0.337	0.165	0.239	documents
	FW14Docs50	0.419	0.292	0.174	0.203	documents, Google API
	FW14Docs75	0.422	0.306	0.106	0.198	documents, Google API
	FW14Search100	0.505	0.425	0.278	0.384	snippets, Google API
	FW14Search50	0.517	0.426	0.271	0.404	snippets, Google API
	FW14Search75	0.461	0.366	0.256	0.345	snippets, Google API
udel	udelftrsbs	0.355	0.272	0.166	0.255	documents
	udelftrssn	0.216	0.174	0.147	0.149	snippets
uiucGSLIS	uiucGSLISf1	0.348	0.249	0.101	0.212	documents
	uiucGSLISf2	0.361	0.274	0.179	0.213	documents
ut	UTTailyG2000	0.323	0.251	0.143	0.224	documents

Table 3: Results for the Resource Selection task.

University of Illinois (uiucGSLIS) [14]

The team from Illinois submitted 2 runs. The first (uiucGSLISf1) ranks resources by their query clarity (defined as the KL-divergence between the query and collection language models). The second (uiucGLSISf2) uses the 'collection frequency - inverse document frequency' score, with slightly better results.

University of Delaware (udel) [1]

The udel team selected resources for a particular query, based on their contribution to those 100 results that were ranked highest according to the query-likelyhood model for the given query. By repeating the experiment based on an index of snippets (with the run udelftrssn), and one based on sampled pages (udelftrsbs), the best performance was reached for the one based on full sampled pages.

University of Stavanger (NTNUiS) [2]

In the previous edition of the track, NTNUiS experimented with two approaches: Collection-Centric and Document-Centric models. This year, they explored learning to rank to combine these strategies. A learning to rank model trained on data from Fedweb'13 (run NTNUiSrs2) performed best. However, a model trained on data from both Fedweb'12 and Fedweb'13 performed worse, achieving even a lower performance than their baseline approach (NTNUiSrs1) that only uses a document-centric model.

University of Twente (ut) [9]

The run UTTailyG2000 was based on the Taily system, originally designed for efficient shard selection for centralized search.

University of Padova (UPD) [6]

Besides vertical selection, the University of Padova also investigated the TWF.IRF scheme for resource selection. They showed that stemming has no significant influence on the effectiveness, whereas stop-word removal does improve the TWF.IRF ranking.

University of Lugano (ULugano) [7]

Their resource selection runs followed approaches that combine relevance and opinion. The relevance of the resource were calculated by the ReDDE resource selection method on the sampled representation of the resources while the opinion mining was based on counting the number of sentiment terms (defined by the external resource SentiWordNet) appearing in documents of each resource. They ultimately submitted three runs, among which ULuganoDFR only utilized a traditional resource selection approach, whereas the other two runs (ULuganoColL2 and ULuganoDocL2) utilized different ways to re-rank based on opinions. However, in the experiments, the opinions do not seem to improve the resource selection performance.

7. RESULTS MERGING

7.1 Evaluation

An important new condition in the Results Merging task, as compared to the analogous FedWeb 2013 task, is the requirement that each Results Merging run had to be based on a particular Resource Selection run. More in particular, only results from the top 20 highest ranked resources in the selection run were allowed in the merging run. Additionally, participants were asked to submit at least one run based on the Resource Selection baseline run provided by the organizers. The evaluation results for the results merging task are shown in Table 4 (runs based on provided baseline) and Table 5 (runs based on participants own resource selection runs), displaying for a number of metrics the average per run over all topics.

Different evaluation measures are shown:

1. nDCG@20 (official RS metric), with the gain of duplicates set to zero (see below), and where the reference covers all results over all resources.
2. nDCG@100: analogous.
3. nDCG@20_dups: analogous to nDCG@20, but without penalizing duplicates.
4. nDCG@20_loc: again an nDCG@20 measure, with duplicate penalty, whereby all results not originating from the top 20 resources of the chosen selection run, are considered non-relevant.
5. nDCG-IA@20: intent-aware nDCG@20 (see [19]), again with duplicate penalty and possibly relevant results from all resources, where each vertical intent is weighted by the corresponding intent probability.

Penalizing duplicates means that after the first occurrence of a particular result in the merged list for a query, all consecutive results that refer to the same web page as that first result, receive the default relevance level of non-relevance. The goal of reporting the nDCG@20_loc measure is to allow comparing reranking strategies only, not influenced by the quality of the corresponding resource selection run, and where an ideal ranking leads to a value of 1. The other reported nDCG@20 values measure the total effectiveness of both the selection and the merging strategies. For ideal ranking, given a selection run, the highest possible value may well be below one, as the denominator can contain contributions from resources outside of the considered 20. The vertical intent probabilities for the nDCG-IA@20 measure are calculated as follows: (i) the quality of each vertical is quantified by the maximum score of the resource the vertical contains, where the score of each resource is measured by the graded precision of the top retrieved documents in the resource, and (ii) the vertical intent probability is obtained by normalizing the vertical score obtained in (i) across all the verticals.

7.2 Analysis

The top 5 performing runs overall are by ICTNET (ICTNETRM06, ICTNETRM07, ICTNETRM04, ICTNETRM05, ICTNETRM03). These runs were based on the official baseline, which the organizers has chosen as ICTNET's ICTNETRS06 run. Interestingly, the higest ranked run ICTNETRM06 (according to the official metric) was obtained by removing duplicates from the already high-scoring run ICTNETRM05, with a resulting increase in nDCG@20 of 5%. Note that the score from ICTNETRM06 according to the official metric remains almost constant, compared to the metric nDCG@20_dups that does include the gain from duplicates, whereas ICTNETRM05 would be rated 14% higher. This

Task 3: Results Merging							
Group ID	Run ID	**nDCG@20**	nDCG@100	nDCG@20_dups	nDCG@20_loc	nDCG@100_loc	nDCG-IA@20
CMU_LTI	googTermWise7	0.286	0.319	0.320	0.395	0.632	0.102
	googUniform7	0.285	0.318	0.322	0.389	0.628	0.101
	plain	0.277	0.316	0.312	0.379	0.623	0.098
	sdm5	0.276	0.315	0.315	0.379	0.623	0.096
ECNUCS	basedef	0.289	0.300	0.336	0.397	0.593	0.095
ICTNET	ICTNETRM01	0.247	0.307	0.361	0.338	0.599	0.080
	ICTNETRM02	0.309	0.305	0.314	0.362	0.512	0.095
	ICTNETRM03	0.348	0.311	0.350	0.405	0.522	0.111
	ICTNETRM04	0.381	0.271	0.386	0.451	0.456	0.121
	ICTNETRM05	0.354	0.354	0.492	0.497	0.706	0.123
	ICTNETRM06	0.402	0.338	0.407	0.473	0.571	0.132
	ICTNETRM07	0.386	0.331	0.390	0.451	0.557	0.123
SCUTKapok	SCUTKapok1	0.313	0.293	0.316	0.367	0.492	0.097
	SCUTKapok2	0.319	0.316	0.361	0.442	0.624	0.106
	SCUTKapok3	0.314	0.294	0.317	0.367	0.491	0.097
	SCUTKapok4	0.318	0.299	0.320	0.370	0.497	0.099
	SCUTKapok5	0.320	0.321	0.344	0.442	0.629	0.102
	SCUTKapok6	0.323	0.298	0.325	0.377	0.497	0.101
	SCUTKapok7	0.322	0.320	0.361	0.446	0.627	0.107
ULugano	ULugFWBsNoOp	0.251	0.296	0.304	0.355	0.588	0.083
	ULugFWBsOp	0.224	0.273	0.271	0.314	0.545	0.072
dragon	FW14basemR	0.322	0.318	0.361	0.446	0.626	0.107
	FW14basemW	0.260	0.298	0.312	0.367	0.592	0.086

Table 4: Results for the Results Merging task based on the official baseline run.

Task 3: Results Merging							
Group ID	Run ID	**nDCG@20**	nDCG@100	nDCG@20_dups	nDCG@20_loc	nDCG@100_loc	nDCG-IA@20
ULugano	ULugDFRNoOp	0.156	0.204	0.157	0.193	0.362	0.035
	ULugDFROp	0.146	0.195	0.149	0.180	0.346	0.033
dragon	drexelRS1mR	0.219	0.298	0.222	0.264	0.491	0.059
	drexelRS4mW	0.144	0.244	0.148	0.177	0.420	0.036
	drexelRS6mR	0.198	0.270	0.194	0.232	0.443	0.050
	drexelRS6mW	0.196	0.270	0.193	0.231	0.444	0.049
	drexelRS7mW	0.250	0.305	0.249	0.318	0.535	0.070

Table 5: Results for the Results Merging task not based on the official baseline run.

confirms the intuitive idea that among the highly relevant (and hence top ranked) results, there are many duplicates (most likely returned by different resources).

The teams SCUTKapok (SCUTKapok6, SCUTKapok7) and dragon (FW14basemR) perform well as well, based on variations on round robin merging, and normalizing document scores based on the resource selection results, respectively.

We further note that the ranking of all submitted runs based on the intent-aware metric nDCG-IA@20 highly correlates with the nDCG@20-based ranking (rank correlation $\rho = 0.95$). Also, despite the clear absolute benefit of removing duplicates (with regard to the official metric nDCG@20), the rank correlation between systems scored on nDCG@20 vs. nDCG@20_dups is high, too ($\rho = 0.89$). The metric nDCG@20_loc, only measuring the reranking capabilities of the proposed methods, independent of the quality of the underlying resource selection baseline, highly correlates with nDCG@20 as well ($\rho = 0.91$). It can also be observed that the correlation when comparing the rank order of runs for nDCG@20 with nDCG@100 is less strong ($\rho = 0.66$).

7.3 Participant Approaches

Chinese Academy of Sciences (ICTNET) [8]

ICTNET proposed various methods for this task, as in the vertical selection and resource selection tasks. Their lowest performant run (ICTNETRM01) is based on IR heuristics, but they also submitted a variant with duplicates filtered out (ICTNETRM02), scoring significantly higher. They again used the resources' pagerank and the LSI model (runs ICTNETRM03 and ICTNETRM04). Their most successful runs however (also the overall best performing runs), were obtained by combining these methods using an ensemble method (ICTNETRM05, ICTNETRM06, ICTNETRM07), whereby the run without duplicates scores best (ICTNETRM06).

South China University of Technology (SCUTKapok) [17]

The team from South China University of Technology has investigated various alterations to the basic round robin method, with significant improvements by taking into account the resource selection baseline, the verticals the resources belong to, and removing duplicates.

Drexel University (dragon) [16]

Their result merging runs were based on normalizing the document score based on the resource score by a simple multiplication. The resource score was determined by the resource selection approach (based on either the raw score or the resource ranking position). On the other side, the document score was based on its reciprocal rank of the selected resource. Ultimately, the rank based resource score combined with the document score on the RS baseline provided by the FedWeb team performed the best (drexelRS7mW).

East China Normal University (ECNUCS) [10]

The ECNUCS results merging run (basedef) simply returns the output of the official FedWeb resource selection baseline.

Carnegie Mellon University (CMU_LTI) [13]

They only participated in the results merging task and submitted several runs based on the baseline. For their baseline run, they used language modeling with Dirichlet smoothing by indexing the search result snippets using the Indri toolkit. In addition, they experimented with a sequential dependence model (sdm5) where the similarity is not only based on individual terms, but also on bigrams (exact match and unordered window). They also explored query expansion using word-vector representations released by Google (googUniform7 and googTermWise7). While the SDM model performed best on the FedWeb13 dataset, the query expansion strategies performed slightly better on the FedWeb14 dataset.

University of Lugano (ULugano) [7]

The four submitted runs were intended to experiment whether diversifying the final merged result list to cover different sentiments, namely positive, negative and neutral, would be helpful. Therefore, both relevance and opinion scores of documents were considered when conducting result merging and a retrieval-interpolated diversification approach was utilized. The differences of the four submitted runs were based on whether they included sentiment diversification or not, and which resource selection baseline they utilized. However, opinion diversification did not boost the performance.

8. CONCLUSIONS

In FedWeb 2014, the second and final edition of the TREC Federated Web Search Track, 12 teams participated in one or more of the challenges Vertical Selection, Resource Selection, and Results Merging, with a total of 106 submitted system runs. We introduced an online evaluation system for system preparations, which turned out a success and in our opinion led to an increased effort into composing well-performing runs. This year's most effective methods are in general more complicated, as compared to the FedWeb 2013 submissions, with the appearance of a number of machine learning methods, besides more traditional information retrieval methods.

We discussed the creation of the FedWeb 2014 dataset and relevance judgments, analyzed the relevance distributions over the test topics and different verticals in our system of 149 online search engines, and for each of the main tasks, listed the performance of the submitted runs, as a set of several evaluation measures. With the individual descriptions of the participants' approaches, this overview paper also provides insights into which methods are best suited for the different tasks.

9. ACKNOWLEDGMENTS

This work was funded by the Folktales As Classifiable Texts (FACT) project in The Netherlands, by the Dutch national project COMMIT, and by Ghent University - iMinds in Belgium.

10. REFERENCES

[1] A. Bah, K. Sabhnani, M. Zengin, and B. Carterette. University of delaware at TREC 2014. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[2] K. Balog. NTNUiS at the TREC 2014 federated web search track. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the*

7th International World Wide Web Conference, WWW '98, 1998.

[4] T. Demeester, R. Aly, D. Hiemstra, D. Nguyen, D. Trieschnigg, and C. Develder. Exploiting user disagreement for web search evaluation: An experimental approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM 2014)*, pages 33–42. ACM, 2014.

[5] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the trec 2013 federated web search track. In *The 22nd Text Retrieval Conference (TREC 2013)*, 2013.

[6] E. Di Buccio and M. Melucci. University of padova at TREC 2014: Federated web search track. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[7] A. Giachanou, I. Markov, and F. Crestani. Opinions in federated search: University of lugano at TREC 2014 federated web search track. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[8] F. Guan, S. Zhang, C. Liu, X. Yu, Y. Liu, and X. Cheng. ICTNET at federated web search track 2014. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[9] D. Hiemstra and R. Aly. U. twente at trec 2014 - two selfless contributions to web search evaluation. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[10] S. Jin and M. Lan. Simple may be best - a simple and effective method for federated web search via search engine impact factor estimation. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[11] J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[12] Y. Li, Z. Zheng, and H. Dai. Kdd cup-2005 report: Facing a great challenge. *SIGKDD Explorations*, 7(2):91–99, 2005.

[13] S. Palakodety and J. Callan. Query transformations for result merging. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[14] G. Sherman, M. Efron, and C. Willis. The university of illinois' graduate school of library and information science at TREC 2014. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[15] Q. Wang, S. Shi, and W. Cao. RUC at TREC 2014: Select resources using topic models. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[16] H. Zhao and X. Hu. Drexel at TREC 2014 federated web search track. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[17] J. Zhou, Y. Xie, S. Dong, and Z. Chen. SCUTKapok at the TREC 2014 federated web task. In *The 23rd Text Retrieval Conference (TREC)*, 2014.

[18] K. Zhou, T. Demeester, D. Nguyen, D. Hiemstra, and D. Trieschnigg. Aligning vertical collection relevance with user intent. In *ACM International Conference on Information and Knowledge Management (CIKM 2014)*, 2014.

[19] K. Zhou, M. Lalmas, T. Sakai, R. Cummins, and J. M. Jose. On the Reliability and Intuitiveness of Aggregated Search Metrics. In *ACM International Conference on Information and Knowledge Management (CIKM 2013)*, 2013.

APPENDIX

A. FEDWEB 2014 SEARCH ENGINES

ID	Name	Vertical	ID	Name	Vertical
e001	arXiv.org	Academic	e100	Chronicling America	News
e002	CCSB	Academic	e101	CNN	News
e003	CERN Documents	Academic	e102	Forbes	News
e004	CiteSeerX	Academic	e104	JSOnline	News
e005	CiteULike	Academic	e106	Slate	News
e007	eScholarship	Academic	e108	The Street	News
e008	KFUPM ePrints	Academic	e109	Washington post	News
e009	MPRA	Academic	e110	HNSearch	Shopping
e010	MS Academic	Academic	e111	Slashdot	News
e011	Nature	Academic	e112	The Register	News
e012	Organic Eprints	Academic	e113	DeviantArt	Photo/Pictures
e013	SpringerLink	Academic	e114	Flickr	Photo/Pictures
e014	U. Twente	Academic	e115	Fotolia	Photo/Pictures
e015	UAB Digital	Academic	e117	Getty Images	Photo/Pictures
e016	UQ eSpace	Academic	e118	IconFinder	Photo/Pictures
e017	PubMed	Academic	e119	NYPL Gallery	Photo/Pictures
e018	LastFM	Audio	e120	OpenClipArt	Photo/Pictures
e019	LYRICSnMUSIC	Audio	e121	Photobucket	Photo/Pictures
e020	Comedy Central	Video	e122	Picasa	Photo/Pictures
e021	Dailymotion	Video	e123	Picsearch	Photo/Pictures
e022	YouTube	Video	e124	Wikimedia	Photo/Pictures
e023	Google Blogs	Blogs	e126	Funny or Die	Video
e024	LinkedIn Blog	Blogs	e127	4Shared	General
e025	Tumblr	Blogs	e128	AllExperts	Q&A
e026	WordPress	Blogs	e129	Answers.com	Q&A
e028	Goodreads	Books	e130	Chacha	Q&A
e029	Google Books	Books	e131	StackOverflow	Q&A
e030	NCSU Library	Academic	e132	Yahoo Answers	Q&A
e032	IMDb	Encyclopedia	e133	MetaOptimize	Q&A
e033	Wikibooks	Encyclopedia	e134	HowStuffWorks	Encyclopedia
e034	Wikipedia	Encyclopedia	e135	AllRecipes	Recipes
e036	Wikispecies	Encyclopedia	e136	Cooking.com	Recipes
e037	Wiktionary	Encyclopedia	e137	Food Network	Recipes
e038	E! Online	Entertainment	e138	Food.com	Recipes
e039	Entertainment Weekly	Entertainment	e139	Meals.com	Recipes
e041	TMZ	Entertainment	e140	Amazon	Shopping
e043	Addicting games	Games	e141	ASOS	Shopping
e044	Amorgames	Games	e142	Craigslist	Shopping
e045	Crazy monkey games	Games	e143	eBay	Shopping
e047	GameNode	Games	e144	Overstock	Shopping
e048	Games.com	Games	e145	Powell's	Shopping
e049	Miniclip	Games	e146	Pronto	Shopping
e050	About.com	Encyclopedia	e147	Target	Shopping
e052	Ask	General	e148	Yahoo! Shopping	Shopping
e055	CMU ClueWeb	General	e152	Myspace	Social
e057	Gigablast	General	e153	Reddit	Social
e062	Baidu	General	e154	Tweepz	Social
e063	Centers for Disease Control and Prevention	Health	e156	Cnet	Software
e064	Family Practice notebook	Health	e157	GitHub	Software
e065	Health Finder	Health	e158	SourceForge	Software
e066	HealthCentral	Health	e159	bleacher report	Sports
e067	HealthLine	Health	e160	ESPN	Sports
e068	Healthlinks.net	Health	e161	Fox Sports	Sports
e070	Mayo Clinic	Health	e163	NHL	Sports
e071	MedicineNet	Health	e164	SB nation	Sports
e072	MedlinePlus	Health	e165	Sporting news	Sports
e075	University of Iowa hospitals and clinics	Health	e166	WWE	Sports
e076	WebMD	Health	e167	Ars Technica	Tech
e077	Glassdoor	Jobs	e168	CNET	Tech
e078	Jobsite	Jobs	e169	Technet	Tech
e079	LinkedIn Jobs	Jobs	e170	Technorati	Tech
e080	Simply Hired	Jobs	e171	TechRepublic	Tech
e081	USAJobs	Jobs	e172	TripAdvisor	Travel
e082	Comedy Central Jokes.com	Jokes	e173	Wiki Travel	Travel
e083	Kickass jokes	Jokes	e174	5min.com	Video
e085	Cartoon Network	Kids	e175	AOL Video	General
e086	Disney Family	Kids	e176	Google Videos	Video
e087	Factmonster	Kids	e178	MeFeedia	Video
e088	Kidrex	Kids	e179	Metacafe	Video
e089	KidsClicks!	Kids	e181	National geographic	General
e090	Nick jr	Kids	e182	Veoh	Video
e092	OER Commons	Encyclopedia	e184	Vimeo	Video
e093	Quintura Kids	Kids	e185	Yahoo Screen	Video
e095	Foursquare	Local	e200	BigWeb	General
e098	BBC	News			

B. FEDWEB 2014 EVALUATION QUERIES

ID	Query
7015	the raven
7044	song of ice and fire
7045	Natural Parks America
7072	price gibson howard roberts custom
7092	How much was a gallon of gas during depression
7111	what is the starting salary for a recruiter
7123	raleigh bike
7137	Cat movies
7146	why do leaves fall
7161	dodge caliber
7167	aluminium extrusion
7173	severed spinal cord
7174	seal team 6
7176	weather in nyc
7185	constitution of italy
7194	hobcaw barony
7197	contraceptive diaphragm
7200	uss stennis
7205	turkey leftover recipes
7207	earthquake
7211	punctuation guide
7212	mud pumps
7215	squamous cell carcinoma
7216	salmonella
7222	route 666
7230	council bluffs
7235	silicone roof coatings
7236	lomustine
7239	roundabout safety
7242	hague convention
7249	largest alligator on record
7250	collagen vascular disease
7252	welch corgi
7261	elvish language
7263	hospital acquired pneumonia
7265	grassland plants
7274	detroit riot
7293	basil recipe
7299	row row row your boat lyrics
7303	what causes itchy feet
7307	causes of the cold war
7320	cayenne pepper plants
7326	volcanoe eruption
7328	reduce acne redness
7431	navalni trial
7441	barcelona real madrid goal messi
7448	running shoes boston
7486	board games teenagers
7491	convert wav mp3 program
7501	criquet miler